Applied Artificial Intelligence

An International Journal

ISSN: 0883-9514 (Print) 1087-6545 (Online) Journal homepage: https://www.tandfonline.com/loi/uaai20

An Algorithmic Scheme for Statistical Thesaurus Construction in a Morphologically Rich Language

Chaya Liebeskind, Ido Dagan & Jonathan Schler

To cite this article: Chaya Liebeskind, Ido Dagan & Jonathan Schler (2019) An Algorithmic Scheme for Statistical Thesaurus Construction in a Morphologically Rich Language, Applied Artificial Intelligence, 33:6, 483-496, DOI: 10.1080/08839514.2019.1583447

To link to this article: https://doi.org/10.1080/08839514.2019.1583447

Published online: 27 Feb 2019.

APPLIED ARTIFICIAL INTELLIGENCE
2019, VOL. 33, NO. 6, 483–496
https://doi.org/10.1080/08839514.2019.1583447

Taylor & Francis
Taylor & Francis Group

An Algorithmic Scheme for Statistical Thesaurus Construction in a Morphologically Rich Language

Chaya Liebeskind, Ido Dagan, and Jonathan Schler

Department of Computer Science, Bar Ilan University, Ramat Gan, Israel

ABSTRACT

Corpus-based automatic thesaurus construction uses linguistic methods, such as Part-of-Speech taggers and parsers, which often perform poorly on MRLs. Therefore, in this paper, we focused on the complex task of adapting corpus-based thesaurus construction methods for MRLs. We investigated two statistical approaches for thesaurus construction; a) a first-order co-occurrence-based approach and b) a second-order distributional-based approach. We explored alternative levels of morphological term representations complemented by grouping the morphological variants. We then introduced and adopted a generic algorithmic scheme for thesaurus construction in MRLs for both first-order and second-order approaches. Our scheme investigated alternative representation levels and offered alternative configurations. We demonstrated the empirical benefits of our methodology for a diachronic Hebrew thesaurus construction. We used morphological analysis tools, defined and applied a new annotation scheme, and demonstrated its optimal configuration, which outperforms the baseline for both first and second order corpus-based thesaurus construction approaches.

Introduction

Corpus-based thesaurus construction is an active research area (Curran and Moens 2002; Kilgarriff 2003; Rapp 2002; Rychlý and Kilgarriff 2007). Whereas most prior work on thesaurus construction focused on English, we are interested in applying these methods to a Morphologically Rich Language (MRL), characterized by highly productive morphological processes (inflection, agglutination, compounding) often produce many word forms for any given root.

Thesauri usually provide related terms for each target term. Since both target and related terms correspond to lemmas, statistical collection is commonly applied at the lemma level, using a morphological analyzer and tagger (Peirsman, Heylen, and Speelman 2008; Rapp 2008). However, due to the rich challenging morphology of MRL, such tools are often limited.

CONTACT Chaya Liebeskind ✉ liebchaya@gmail.com Department of Computer Science, Jerusalem College of Technology, Lev Academic Center, Jerusalem, 9116001, Israel

Therefore, we propose several alternative methods for generating a corpus-based thesaurus in MRL. In particular, we propose three options for term representation, surface form (e.g. telling), best lemma (tell) and multiple lemmas (tell (VBG), telling (JJ)), supplemented with grouping of morphological term variants.

We demonstrate the empirical benefit of our methodology for a diachronic Hebrew thesaurus construction and show that exploring the alternative morphological term representation levels in statistical thesaurus construction is useful for optimizing thesaurus precision and coverage.

This paper follows up on our earlier short paper (Liebeskind, Dagan, and Schler 2012), and extends that work substantially.

Background

In the last decades, typically, two statistical approaches for identifying semantic relationships between words were investigated: first-order: co-occurrence-based methods and second-order: distributional similarity methods.

First-Order Co-Occurrence-Based Thesauri

Co-occurrence data for thesaurus construction presumes that words that tend to occur together in the same context are likely to have similar or related meanings (Qiu and Frei 1993). Common metrics for measuring co-occurrence strength are the Dice coefficient (Smadja, McKeown, and Hatzivassiloglou 1996) and the Pointwise Mutual Information (PMI) (Church and Hanks 1990). Both first-order metrics consider the number of times each candidate term co-occurs with the target term in the same context, relative to their total frequencies in the corpus.

Second-Order Distributional-Based Thesauri

Over the last decade, distributional thesauri have attracted much interest (Bhagat, Pantel, and Hovy 2007; Erk and Padó 2008; Pantel and Ravichandran 2004; Weeds and Weir 2003). Distributional thesauri are based on the Distributional Similarity Hypothesis, which suggests that words that occur within similar contexts are semantically similar (Harris 1968).

First, by collecting context words, a feature vector is generated for each word. Each vector entry represents a type of occurrence relation, usually a co-occurring word and possibly including the syntactic relation between the two co-occurring words (Cimiano, Hotho, and Staab 2005; Erk and Padó 2008; Lee 1999; Lin 1998). Each feature is assigned a weight indicating its “relevance” (or association) to the given word.

Given the representation of words as context vectors, a second-order measure for the degree of similarity between pairs of vectors is needed. Among the similarity measures that were used are: Jaccard's coefficient (Gasperin et al. 2001), Cosine-similarity (Caraballo 1999; Pantel and Ravichandran 2004) and Lin's mutual information metric (Lin 1998).

Methodology and Algorithm

In this paper, we assume that the list of target terms is given as input. We focus on the process of extracting a ranked list of candidate related terms (termed candidate terms) for each target term. The top ranked candidates may be either pruned by threshold, or, more likely, as in our case, presented to a lexicographer for manual filtering.

Roles of Terms in Statistical Thesaurus Construction

Terms in statistical thesaurus construction have different roles. We distinguish between four term roles: target term, related term, candidate term and feature term.

The input term for the statistical construction process is a thesaurus entry (target term). The final output of the statistical extraction process is a related terms list for the thesaurus entry. The statistical extraction, either by first-order or by second-order similarity methods, generates a ranked list of candidate related terms (candidate terms). Often, the top ranked candidates of this list would be considered as the related terms for the thesaurus' target term entry.

Second order methods involve an additional type of term for feature representation. In feature representation, a feature vector is constructed for each term in the corpus by collecting context terms as features. Each such term (feature term) is assigned a weight indicating its association to the given term. Then, second order similarity is calculated between the target term and all the other terms in the corpus. Therefore, each term in the corpus is a potential candidate term. Yet, the ranked list of terms is considered as the final candidate terms list. Although the feature term weight is calculated by first-order similarity, as done for candidate terms in the first-order methods, their roles in statistical thesaurus construction is different.

Term Representation

The statistical extraction process is affected by term representation in the corpus. Usually, both target and related terms in a thesaurus are represented by lemmas, which can be identified by morphological disambiguation tools.

However, we present two alternative approaches for term representation which are less dependent on morphological processing.

Typically, a morphological analyzer produces all possible analyses for a given token in the corpus. Then, a Part Of Speech (POS) tagger selects the most probable analysis and solves morphology disambiguation. However, considering the poor performance of the POS tagger on certain MRLs corpora, we distinguish between these two analysis levels. Consequently, we examined three levels of term representation:

(1) Surface form (surface)
(2) Best lemma, as identified by a POS tagger (best)
(3) All possible lemmas, produced by a morphological analyzer (all)

When the target term is represented in its all lemma representation, we consider each appearance of any possible lemma as an appearance of the target term. However, when a feature or candidate term is represented in its all lemma representation, we assume that the right lemma would accumulate enough statistical prominence throughout the corpus. Therefore, each lemma is considered as a different feature or candidate during the statistics collection process.

The feature and candidate's best lemma representation is context dependent and chosen using a tagger. The best lemma representation of the target term is context-free, since we do not assume any prior knowledge of the target term's context.

Candidate Generation

We applied our methodological scheme for exploring alternative term representations in statistical thesaurus construction for both first-order and second-order statistical extraction. The exploration of the different target term representations mainly aims to analyze the impact of the representation on the amount and quality of the identified target term occurrences in the statistical extraction process. On the other hand, the exploration of the candidate and feature term representations mainly aims to investigate the performance of the similarity measure over the different representations.

First-Order Candidate Extraction

We used the following algorithmic scheme for first-order thesaurus construction. Our input is a target term in one of the possible term representations (surface, best or all). For each target term we retrieved all the contexts in the corpus where the target term appears (in its current form). Then, we defined a set of candidate terms that consisted of all the terms that appear in all these contexts (this again for each of the three possible term representations). Next, a co-occurrence score between the target term and each of the candidates was calculated. Then, candidates were sorted, and the highest rated candidate terms were grouped into lemma-oriented groups. Finally, we ranked the groups by their members' co-occurrence scores and the highest rated groups were considered as the related terms in the thesaurus.

The two choices for target term and candidate representations, which may be either surface, best or all, are independent, resulting in nine possible configurations of the algorithm for representing both the target term and the candidate terms. Thus, these 9 configurations cover the space of possibilities for term representation. Exploring all of them in a systematic manner should reveal the best configuration in a particular setting.

Second-Order Candidate Extraction

For second-order thesaurus construction, we constructed feature vectors for both the target term and the candidates and compared them by distributional similarity measures. Since every word in the corpus was a potential candidate for sharing a common context with the target term, we constructed a feature vector for each word in the corpus. The feature extraction process for a target term and a candidate term was identical. First, we represented the term in one of the three term representations. Then, for each term we retrieved the contexts in the corpus where it appeared and defined a set of feature terms consisting of all the terms in all these contexts. Finally, a co-occurrence score between the term and each of the features was calculated and the term's feature vector was stored. After constructing feature vectors for all candidates, we scored candidates by their vector similarity with the target term vector, candidates were sorted, and the highest rated candidate terms were grouped into lemma-oriented groups. Finally, we ranked the groups by their members' distributional similarity scores and the highest rated groups became related terms in the thesaurus.

There are three independent choices for target term, feature and candidate representations, resulting in 27 configurations which cover the range of possibilities for term representation in second-order (distributional similarity) thesaurus.

Grouping

Both algorithms for candidate extraction suggest grouping the extracted candidates before considering them for the thesaurus. Grouping aims at bringing related terms with the same lemma into groups by morphological tools. Obviously, grouping is mostly needed for surface-level representation to group different inflections of the same lemma. Yet, because morphological tools often assign incorrect lemma, we note that grouping was also found beneficial for the lemma-level representations. Moreover, the tagger often identifies slightly different lemmas for the same term. For example, a title tag is sometimes assigned to a noun prefixed with a determiner, and as a result the term h-šwpT[1] (the judge) became a separate lemma instead of conflating it with the lemma šwpT (judge, without the determiner). Therefore, the grouping process is beneficial also for grouping these two lemma variants into a single group.

We grouped term variants together based on their most probable lemma. Each candidate term's most probable lemma was selected independently and terms were grouped together if their most probable lemmas were equal. The probability of a lemma was calculated by a context-free method for acquiring morpholexical probabilities from an untagged corpus (Levinger, Itai, and Ornan 1995).

After applying the grouping algorithm, we re-ranked the groups with the best at the top of the list. We investigated three scoring approaches for group ranking; maximization, averaging, and summation. In addition, we investigated a unification approach which calculates the dice coefficient between the union of all group members' occurrences and the target term.

The contribution of this paper is a comprehensive scheme for experimenting with different possible term representations, specifically geared for MRLs. Each corpus and language-specific tool set might yield a different optimal configuration.

Application to Historical Hebrew Jewish Corpus

The Responsa Corpus

Our research focused on the construction of a diachronic thesaurus for the Responsa project.[2] The corpus includes questions on many daily issues including law, health, commerce, marriage, education, and Jewish customs posed to rabbis along with their detailed rabbinic answers (each question and answer in a separate article). It contains 76,710 articles and about 100 million word tokens, and was used for previous IR and NLP research (HaCohen-Kerner, Kass, and Peretz 2008; Koppel 2011; Zohar et al. 2013).

Our corpus represents more than a thousand years of various genres and styles of world-wide Jewish literary creativity. Responsa present juristic negotiation with arguments citing earlier sources, such as the Talmud and its commentators, legal codes, and earlier response (Koppel 2011) and pose an increased challenge for thesaurus developers.

Recently, Zohar et al. (2013) investigated methods for automatic thesaurus construction in relation to the Responsa Corpus. They applied both co-occurrence and distributional similarity approaches and suggested a unified algorithm. However, Zohar et al. (2013) worked at the surface word level without dealing with morphology. Since they observed many inflections of the same lemma in the constructed thesaurus, in the evaluation phase, they classified each related term to a group based on its lemma.

Thesaurus Goal

The goal of our thesaurus is to bridge the gap between modern and ancient language in the domain of rabbinic/Jewish literature. The potential user would be able to search our thesaurus for a modern term and get related terms from previous periods.

The cross-period thesaurus for the Responsa Project was aimed to be comprehensive. Therefore, we extracted two types of semantic relations: expansion relation and association relation. The expansion relation requires a meaning correlation, including synonyms, hyponyms, and antonyms. In contrast, the association relation demands a common subject or context. For example, an expansion for the target term apilpsih (epilepsy) is the synonym mxlt hnpilh (literally "falling disease") and an associated related term is prkws (convulsion). Since we do not distinguish between the two types of semantic relations, we actually search for any topical similarity between the target term and its related terms. The decision of which relation to include in the final thesaurus was made by a lexicographer.

Target Terms Collection

Since the outcome of this research is a thesaurus for the Responsa Project, we needed an input set of target terms for inclusion. Ideally, the target terms list should consist of thousands of terms, whose collection is not a trivial task. We examined two publicly available key-lists, the University of Haifa's entry list[3] and Hebrew Wikipedia entries.[4] However, many of these entries, such as person names and place names were not relevant as target terms for this corpus. Therefore, we first filtered them by Hebrew Named Entity Recognition (NER).[5] Then, we built a target term list from the intersection of the filtered key-lists. Finally, we manually filtered additional irrelevant terms and constructed an appropriate target terms list of 5000 terms. The results reported in this paper were obtained from a sample of 108 randomly selected terms from this target terms list.

Morphological Tools

Unfortunately, due to the different genres in the Responsa corpus, available tools for Hebrew processing perform poorly on this corpus. In a preliminary experiment, the POS tagger (Adler and Elhadad 2006) accuracy on the Response Corpus was less than 60%, while the accuracy of the same tagger on modern Hebrew corpora is 90% (Bar-haim, Sima'an, and Winter 2008).

Therefore, we preferred token-based lemmatization over POS lemmatization. Token-based lemmatization assigns the lemma token while ignoring the POS. For example, the lemma of the Hebrew token spr is spr. Even thought, it can be analyzed as either a noun: barber or book, or as a verb: counted or told. In addition, token-based lemmatization is a suitable solution for the Niqud-less Responsa project. We assume that a lexicographer will distinguish between semantic differences.

For this project, we used the MILA Hebrew Morphological Analyzer (Itai and Wintner 2008; Yona and Wintner 2008) and the (Adler and Elhadad 2006) POS tagger for lemma representation. The latter had two important characteristics: First, flexibility – this tagger allows adapting the estimates of the prior (context-independent) probability of each morphological analysis in an unsupervised manner, from an unlabelled corpus of the target domain (Goldberg, Adler, and Elhadad 2008). The second advantage is its mechanism for analyzing unknown tokens (Adler et al. 2008). Since about 50% of the words in our corpora are unknown (with respect to MILA's lexicon), such mechanism was essential.

Evaluation Setting

Since manual annotation is expensive and time consuming, we constructed a gold-standard by annotating for each configuration the top 15 groups constructed from the top 50 candidate terms, for each target term.[6] First, the annotator judged the group members. Then, each group with at least one positive term was considered as relevant.

We evaluated the inter-annotator agreement over 200 candidate terms that were randomly sampled from our configurations' output. We annotated them according to the annotation guidelines by two annotators. We observed a Kappa value of 0.73, which is considered as substantial (Landis and Koch 1977).

In our experiments, we compared the performance of our algorithms by four common IR measures: average precision (AP), F1, precision (P), and recall (R). In our case, with no pre-defined thesaurus, we evaluated relative-recall. Our relative-recall considered the related terms from all our automatic related terms extractions. The scores were macro-averaged.

Results

In this section we present the results of the statistical extraction by first-order and second-order extraction methods. The relative-recall is calculated separately for each extraction method. Thus, the gold standard of the first-order method is the groups constructed by the unification of all the first-order configurations. While the gold standard of the second-order method is the groups constructed by the unification of all the second-order configurations. This separation of the methods' gold standards enabled us to better understand the overlap degree of different configuration of the same method.

For statistical extraction, we used Lucene.[7] We took the top-1000 documents retrieved for the target term and extracted either candidate terms or feature terms from them, depending on the similarity measure's type. We considered a document as the context of a term, since we extracted topically related terms and each document of the corpus includes a question on a different topic.

The Dice coefficient was used as our co-occurrence measure for both first-order similarity and feature weighting in second-order similarity. Since we observed that the most informative features received the highest weights and the feature vectors had a long noisy tail, we used vectors of the top-1000 features. At the end of the extraction process, we grouped term variants. Groups were ranked based on the summation approach, with yielded similar results to the maximization approach, but was more effective than the unification and averaging approaches.

First-Order Results

Table 1 compares the performance of all nine term representation configurations. Due to data sparseness, the lemma-based representations of the target term outperformed its surface representation. However, the best results were obtained from surface level candidate representation, which was complemented by grouping term variants to lemmas in the grouping phase.

Furthermore, we note that lemma-based target term representation (best or all) yielded the best R and AP scores, which we consider as most important for the thesaurus construction setting. The improvement over the common default best lemma representation, for both target and candidate, is notable (8 points) and is statistically significant according to the one-sided Wilcoxon signed-rank test (Wilcoxon 1945) at the 0.01 level for both AP and R.

Second-Order Results

We compared the performance of all 27 term representation configurations for the second-order statistical extraction method.

Table 1. Results for first-order method.

Candidate►				
Target▼		Surface	Best	All
Surface	R	39.47	32.3	29.59
	P	24.69	21.29	19.18
	F1	30.38	25.67	23.28
	AP	23.6	17.71	16.22
Best	R	**50.3**	42.12	41.53
	P	**26.48**	23.71	22.45
	F1	**34.69**	30.34	29.15
	AP	**29.41**	23.13	20.32
All	R	49.89	44.8	44.28
	P	24.07	24.01	21.79
	F1	32.48	31.27	29.21
	AP	29.25	24.08	21.88

In general, second-order methods have lower performance than first-order methods on our corpus (Zohar et al. 2013). Although we did not set a frequency threshold on the target terms appearances in the corpus, the first-order method succeeded to retrieve a few related terms for target terms with low frequency. However, the candidate lists of these terms were noisy and often constructed feature vectors of low quality. As a consequence, the results of the second-order method were low.

Furthermore, first-order computation is a mixture of both syntagmatic and paradigmatic associations (Rapp 2002). Since topical similarity, which was targeted in this use case, includes both types of associations, it suggests why in the current case study the first-order methods outperformed the second-order methods.

Table 2 presents the performance of all the 27 configurations. The best results of the second-order method were obtained from target term representation at the all-lemma level, feature representation at the surface level and candidate representation at the surface level, which was complemented by grouping term variants to lemmas in the grouping phase. This configuration yielded the best R and P scores and its improvement over the common default best lemma representation is notable (12 and 5 points respectively) and is statistically significant according to the one-sided Wilcoxon signed-rank test (Wilcoxon 1945) at the 0.01 level for R and AP, and at level 0.05 for P. The default best lemma representation has relatively low performance. It is ranked 19 out of the 27 configurations, while, even the default surface representation for both target, feature and candidate has a higher rank (10).

The best AP of the second-order method was obtained from target term representation at the best lemma level, feature representation at the surface level and candidate representation at the surface level.

Due to data sparseness, the recall of the lemma-based representations of the target term outperform its surface representation.

Table 2. Results for second-order all 27 term representations.

Target►		All Lemma			Best lemma			Surface		
Candidate►		Surface	Best	All	Surface	Best	All	Surface	Best	All
Feature▼	R	**34.56**	32.45	31.44	33.3	29.81	28.42	26.03	23.43	22.21
Surface	P	**15.25**	14.75	14.26	14.91	14.21	13.58	15.23	13.88	13.13
	F1	**21.16**	20.28	19.62	20.59	19.25	18.38	19.22	17.43	16.5
	AP	16.52	16.1	15.21	**16.78**	14.45	14.15	13.18	11.34	10.3
Best	R	28.77	24.79	22.9	24.86	*22.76*	22.93	21.89	18.8	17.82
	P	12.22	11.3	10.68	11.7	*10.57*	10.69	12.72	10.34	10.34
	F1	17.16	15.52	14.57	15.91	*14.43*	14.58	16.09	13.34	13.09
	AP	14.11	10.9	10.39	12.68	*10.39*	10.27	10.81	7.75	7.79
All	R	30.04	28.01	25.63	24.15	22.77	21.35	22.4	18.56	17.99
	P	12.84	11.85	11.36	11.89	10.57	10.13	13.27	10.75	10.61
	F1	17.99	16.66	15.74	15.93	14.43	13.74	16.66	13.61	13.35
	AP	14.06	12.4	10.27	12.1	10.12	9.16	10.75	8.21	7.74

We further analyze which dimension of representation (target, feature or candidate) has the most impact on the performance's improvement of the best configuration over the default representation of best-lemma representation for all the dimensions (best-best-best[8]). We performed an ablation test, comparing the best performing configuration (all-surface-surface) to three configurations. In the first configuration the target term is represented at the best-lemma default representation level and the two other dimensions are represented by their optimal representation, as observed by the best configuration (best-surface-surface). In the second configuration the features are represented at the best-lemma default representation level and the two other dimensions are represented by their optimal representation (all-best-surface). While in the third configuration the candidates are represented at the best-lemma default representation level and the two other dimensions are represented by their optimal representation (all-surface-best). Table 3 shows the ablation test with the recall's p-value of the hypothesis that the best configuration outperforms the ablation case (one-sided Wilcoxom singed-rank test). For completeness, we also present the default configuration of surface representation for both target, feature and candidate (surface-surface-surface), along with the best lemma default representation for the three dimensions (best-best-best).

The all-best-surface configuration has the lowest recall p-value. The default representation for the feature level decreased the recall more significantly than any of the other ablation cases. Thus, the decision at which level to represent the feature was the most important decision. In addition, even though the best configuration outperforms both default configurations, the statistical significant is higher for the default configuration that utilizes the optimal feature representation (surface-surface-surface). Since the core of the second-order method is the feature vectors' comparison, these findings are not surprising.

The goal of features' lemmatization was to unify derivations and enable the second-order vectors' comparison to recognize an overlap of different derivations of the same term. However, since the lemmatizer has poor performance on our corpus, wrong lemmas were assigned and feature representation at the lemma level was noisy. Thus, the best level for feature representation is the surface level. Related terms share enough common derivation of features and representing features at the surface level well captures their similarity.

Table 3. Ablation test for the best configuration.

Configurations	R	P	F1	AP	P-value
All-surface-surface	34.56	15.25	21.16	16.52	-
Best-surface-surface	33.3	14.91	20.5	16.78	0.1922
All-best-surface	28.77	12.22	17.16	14.11	0.0179
All-surface-best	32.45	14.75	20.28	16.1	0.2119
Best-best-best	22.76	10.57	14.43	10.39	0.0023

Table 4. Error analysis.

Error type	First-order	Second-order	Example
Similar lemma: an incorrect matching between the target term lemma and its corpus entry	15.54	15.04	Target term: xwp (shore) Related terms: xwph (Jewish marriage ceremony) Xpiph (washing hair)
Rare context: the target term lemma appears in a specific context, which is not the target term's representative context	8.06	– – -	Target term: adwmiwt (Edomite) Related term: Dwag (a person with a similar nickname)
Broad context: terms that share a broader context with the target term	62.69	74.8	Target term: alwminiwm (aluminium) Related terms: mcwpin (covered), m`Tpin (wrapped)
Rare terms: infrequent terms with high scores since their few occurrences appear with the target term	16.72	10.16	Target term: abTxh (security) Related term: bwTšqab (a village in Hungary)

Our results show that, in our corpus, using morphology tools for the target term representation contributes to the statistical extraction process. However, they are not essential for feature representation or for candidate representation.

Error Analysis

We performed an error analysis for the best configuration of both the first-order and the second-order methods. We investigated the false-positive related terms for each method in order to identify several reasons for retrieving negative related terms. Since we allowed the annotator to include related terms which hold the association relation with the target term (most of the errors were non-related terms, which were arbitrarily extracted and did not share any context with the target term (69.98% from the first-order method's errors and 82.42% from the second-order method's errors). We further analysis the remaining errors in Table 4.

Conclusions

The primary contribution of this research, concerning the acquisition of a thesaurus for MRLs, is the adaptation of statistical thesaurus construction methods at the morphological level. We presented a methodological scheme for exploring alternative term representations in statistical thesaurus construction for MRL, complemented by lemma-oriented grouping at the end of the process. Our methodological scheme was adopted to first-order co-occurrence based methods and to second-order distributional similarity methods.

We investigated the scheme for a Hebrew cross-period corpus and showed that solving morphological disambiguation "in retrospect" outperformed the default representation approach for non-MRLs. The scheme can be generically applied in other settings. Since we believe a comprehensive thesaurus incorporates Multi Word Expressions (MWE), we suggest extending our methods to include MWE.

Notes

1. To facilitate readability, we use a transliteration of Hebrew using Roman characters; the letters used, in Hebrew lexico-graphic order, are abgdhwzxTiklmns`pcqršt.
2. Corpus kindly provided – http://www.biu.ac.il/jh/Responsa.
3. http://lib.haifa.ac.il/systems/ihp.html.
4. http://he.wikipedia.org.
5. http://www.cs.bgu.ac.il/nlpproj/hebrewNER.
6. Any of the configurations returned at least 15 groups for each of the target terms.
7. http://lucene.apache.org.
8. The format of the configuration notation is: target term representation – feature representation -candidate representation.

References

Adler, M., and M. Elhadad. 2006. An unsupervised morpheme-based Hmm for hebrew morphological disambiguation. In *COLING-ACL*.

Adler, M., Y. Goldberg, D. Gabay, and M. Elhadad. 2008. Unsupervised lexicon-based resolution of unknown words for full morphololological analysis. In *ACL*.

Bar-Haim, R., K. Sima'an, and Y. Winter. 2008, April. Part-of-speech tagging of modern hebrew text. *Natural Language Engineering* 14(2):223–51. doi: 10.1017/S135132490700455X.

Bhagat, R., P. Pantel, and E. Hovy. 2007. LEDIR: An unsupervised algorithm for learning directionality of inference rules. In *EMNLP-CoNLL*.

Caraballo, S. A. 1999. Automatic construction of a hypernym-labeled noun hierarchy from text. In *ACL*.

Church, K. W., and P. Hanks. 1990, March. Word association norms, mutual information, and lexicography. *Computational Linguistics* 16(1):22–29.

Cimiano, P., A. Hotho, and S. Staab. 2005, August. Learning concept hierarchies from text corpora using formal concept analysis. *Journal of Artificial Intelligence Research (JAIR)* 24 (1):305–39. doi: 10.1613/jair.1648.

Curran, J. R., and M. Moens. 2002. Improvements in automatic thesaurus extraction. In *ACL Workshop on Unsupervised Lexical Acquisition*, 59–66. Stroudsburg, PA. doi:10.1044/1059-0889(2002/er01)

Erk, K., and S. Padó. 2008. A structured vector space model for word meaning in context. In *EMNLP*, 897–906. Stroudsburg, PA, USA.

Gasperin, C., P. Gamallo, A. Agustini, G. Lopes, and V. de Lima. 2001. Using syntactic contexts for measuring word similarity. In *The Workshop on Semantic Knowledge Acquisition and Categorisation (ESSLI 2001).*

Goldberg, Y., M. Adler, and M. Elhadad. 2008. Em can find pretty good Hmm pos-taggers (when given a good start. In *ACL*, 746–54.

HaCohen-Kerner, Y., A. Kass, and A. Peretz. 2008. Combined one sense disambiguation of abbreviations. In *ACL-HLT: Short Papers*, 61–64. Stroudsburg, PA.

Harris, Z. S. 1968. *Mathematical Structures of Language*. New York, NY, USA: Wiley.

Itai, A., and S. Wintner. 2008. Language resources for hebrew. *Language Resources and Evaluation* 42 (1):75–98. doi:10.1007/s10579-007-9050-8.

Kilgarriff, A. 2003. Thesauruses for natural language processing. In *Proceedings of the Joint Conference on Natural Language Processing and Knowledge Engineering*, 5–13. Beijing, China.

Koppel, M. 2011. The responsa project: somepromising future directions. In *Language, Culture, Computation. Computing of the Humanities, Law, and Narratives* (pp. 1-8). Springer, Berlin, Heidelberg.

Landis, J. R., and G. G. Koch. 1977. The measurement of observer agreement for categorical data. *Biometrics* 33 (1):159–174. doi:10.2307/2529310.

Lee, L. 1999. Measures of distributional similarity. *ACL*: 25–32.

Levinger, M., A. Itai, and U. Ornan. 1995, September. Learning morpho-lexical probabilities from an untagged corpus with an application to hebrew. *Computational Linguistics* 21 (3):383–404.

Liebeskind, C., I. Dagan, and J. Schler. 2012. Statistical thesaurus construction for a morphologically rich language. In **SEM 2012: The First Joint Conference on Lexical and Computational Semantics*, 59–64. Montréal, Canada. doi:10.1094/PDIS-11-11-0999-PDN

Lin, D. 1998. Automatic retrieval and clustering of similar words. In *COLING-ACL*.

Pantel, P., and D. Ravichandran. 2004. Automatically labeling semantic classes. In *HLT-NAACL*, 321–28.

Peirsman, Y., K. Heylen, and D. Speelman. 2008. Putting things in order. First and second order context models for the calculation of semantic similarity. In *9es Journées Internationales d'Analyse Statistique Des Données Textuelles (JADT)*, Lyon, France.

Qiu, Y., and H.-P. Frei. 1993. Concept based query expansion. In *Proceedings of the 16th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, 160–69. New York, NY, USA: ACM.

Rapp, R. 2002. The computation of word associations: Comparing syntagmatic and paradigmatic approaches.

Rapp, R. 2008. The automatic generation of thesauri of related words for English, French, German, and Russian. *International Journal of Speech Technology* 11 (3–4):147–56. doi:10.1007/s10772-009-9043-7.

Rychlý, P., and A. Kilgarriff. 2007. An efficient algorithm for building a distributional thesaurus (and other sketch engine developments). In *ACL Interactive Poster and Demonstration Sessions*, 41–44. Stroudsburg, PA, USA.

Smadja, F., K. R. McKeown, and V. Hatzivassiloglou. 1996. Translating collocations for bilingual lexicons: A statistical approach. *Computational Linguistics* 22 (1):1–38.

Weeds, J., and D. Weir. 2003. A general framework for distributional similarity. In *EMNLP*.

Wilcoxon, F. 1945. Individual comparisons by ranking methods. *Biometrics Bulletin* 1 (6):80–83. doi:10.2307/3001968.

Yona, S., and S. Wintner. 2008. A finite-state morphological grammar of hebrew. *Natural Language Engineering* 14 (2):173–90. April. doi:10.1017/S1351324906004384.

Zohar, H., C. Liebeskind, J. Schler, and I. Dagan. 2013, April. Automatic Thesaurus Construction for Cross Generation Corpus. *Journal on Computing and Cultural Heritage (JOCCH)* 6(1): 4:1–4 19.